Proceedings of the Design Society, Volume 6: DESIGN 2026
https://doi.org/10.1017/pds.2026.10460

A novel approach to design creativity assessment: a comparative evaluation

Zhengya Gong[1,✉], Ummi Latif[1], Milene Gonçalves[2] and Georgi V. Georgiev[3,1]

[1] Center for Ubiquitous Computing, University of Oulu, Finland, [2] Department of Design, Organisation and Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, The Netherlands, [3] Japan Advanced Institute of Science and Technology, Japan

✉ zhengya.gong@oulu.fi

ABSTRACT: Various methods are offered to measure design creativity. Nonetheless, a debate continues over the development of a trustworthy and objective method to mitigate the influences on the evaluation. Therefore, we propose and develop a novel tool and approach for evaluating design creativity, which reframes creativity evaluation as a structured and data-driven process. This tool represents a significant advancement in the evaluation of design creativity by mitigating objective factors. Importantly, our tool enables researchers and educators to adopt it freely and objectively for evaluation.

KEYWORDS: design creativity, design evaluation, tournament, evaluation tool, comparative judgement

1. Introduction

Design creativity is fundamental to innovation in various fields, such as product design, architecture, engineering, and human-computer interface (Amabile, 1982; Rhodes, 1961). It relates to the capacity of creating ideas that are both novel and useful, hence fostering technical and societal advancement (Abraham, 2022; Bastian et al., 2025; Casakin & Georgiev, 2021; Çelik, 2025). Evaluating design creativity is then crucial in the design field, not only to better understand creative processes conceptually but also to select ideas for production, and foster innovation (Amabile, 1982; Gong, 2024). Nonetheless, evaluating design creativity continues to be a challenge because of its subjective, multifaceted, and context-dependent characteristics (Amabile, 1982; Kudrowitz et al., 2012; Kwon & Kudrowitz, 2019; Niu & Sternberg, 2001). Conventional evaluation methods, such as the Consensual Assessment Technique (CAT) (Amabile, 1982), depend on experts' evaluations of creativity. Although prevalent, CAT (and other creativity metrics) is frequently impacted by several recognized factors, including evaluator bias, order effects, preference inclinations, and aesthetic influences (Kudrowitz et al., 2012; Kwon & Kudrowitz, 2019; Niu & Sternberg, 2001). These factors highlight the need for more scientific, transparent, and reproducible methods to evaluate design creativity.

To address these challenges, we summarized the factors affecting creativity evaluation, which we named **four elements of creativity evaluation (4ECE)**. These include four main elements involved in the evaluation: **evaluator, evaluated product, evaluation process, and external factor**, which are categorized into subjective and objective factors. While subjective factors, such as prior experiences or individual preferences, are difficult to eliminate, specific objective factors might be mitigated through improved evaluation tools and methods. Thus, we developed an innovative **evaluation tool and approach** that implements design creativity evaluation via triadic comparison instead of absolute grading. In this way, evaluators are presented with three ideas simultaneously and instructed to rank them based on a specified criterion, rather than give a score in relation to the entire pool of ideas. The tool generates a comprehensive rating of all ideas through **iterative comparisons** and **computational aggregation**, which has the potential to enhance the reliability and validity of design creativity evaluation.

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

2. Background theory

To better understand the theoretical background behind the current study, we initially clarified the definition and essential items, such as design creativity and design creativity evaluation techniques. In addition, we explained the elements affecting design creativity and summarized them into a framework. Moreover, we highlighted the research gap that inspired and triggered the current study.

2.1. Design creativity

Design creativity refers to the process of generating ideas (e.g., solutions or products) that are both novel and useful within a given design context (Chakrabarti, 2009; Gong et al., 2024). It is a fundamental element in design fields, driving innovation and problem-solving (Abraham, 2022; Bastian et al., 2025; Casakin & Georgiev, 2021; Çelik, 2025). The process often entails iterative cycles of divergent and convergent thinking, when designers initially investigate a broad spectrum of options before assessing or refining chosen ideas (Gonçalves et al., 2016; Gonçalves & Cash, 2021). Design creativity is frequently evaluated based on novelty and usefulness, two criteria that collectively characterize design creativity (Dean et al., 2006; Shah et al., 2003). However, other criteria are often also considered depending on the researcher's focus, such as workability, aesthetic quality and elaboration (Georgiev & Casakin, 2019; Gong et al., 2022, 2024; Sääksjärvi & Gonçalves, 2018). Understanding and evaluating design creativity outcomes is essential to advance the study of design creativity (e.g., what supports it or hinders it) and for developing educational strategies and design-support tools that lead to innovation.

2.2. Evaluation of design creativity

Evaluating design creativity remains a persistent challenge because of its inherently subjective and multidimensional nature (Baer & McKool, 2009; Kaufman et al., 2008). One widely accepted approach is the Consensual Assessment Technique (CAT), positing that creativity should be evaluated by domain experts (Amabile, 1982) on various criteria, such as novelty, usefulness, and overall creativity, frequently employing Likert scales. Other studies utilize metrics to evaluate creativity based on divergent thinking research, encompassing fluency (quantity of ideas) (Verhaegen et al., 2013), flexibility (variety of categories) (Borgianni et al., 2020; Jagtap et al., 2015), and originality (Borgianni et al., 2020; Dean et al., 2006; Shah et al., 2003). Nonetheless, the evaluation method may exhibit inconsistency among evaluators and over time.

2.3. Four elements influencing design creativity evaluation

The evaluation of design creativity is influenced by various factors, with most of the previous studies focusing on methods, evaluators and the quality of sketches of ideas (Amabile, 1982; Christiaans, 2002; Kudrowitz et al., 2012; Niu & Sternberg, 2001). However, prior studies may have overlooked "**external factor**," such as the physical environment, which might potentially impact the evaluation due to the "Context effect" (Schwarz & Sudman, 2012). This refers to the effect of environmental factors on how stimuli are perceived, illustrating that our context significantly influences perception of our daily experiences (Schwarz & Sudman, 2012). Therefore, this study conceptualizes a framework for understanding creativity evaluation, referred to here as the **four elements of creativity evaluation (4ECE)**, including evaluator, evaluation process, evaluated product, and external factor (Figure 1). Among these, the **evaluator** plays a central role: their cognitive abilities, personality traits, preferences, cultural background, and prior experience substantially shape how creativity is evaluated. It is evident that the "**evaluated product**," such as the idea, solution, or product, requires evaluation, and the quality, content, and creativity of them influence the evaluation of design creativity. Furthermore, the "**evaluation process**," specifically the methodology and evaluation setting employed to evaluate design creativity, subsequently influences the evaluation of design creativity. "**External factor**" refers to contextual conditions surrounding the evaluation situation that are not inherent to the evaluator, the evaluated product, or the evaluation process itself, such as the physical environment surrounding the evaluator, which may affect evaluators' perceptions and judgments.

The four elements were categorized into subjective and objective factors, illustrated in Figure 1 by orange (subjective factor) and blue (objective factors) backgrounds. The categorization is based on their primary source and the degree to which each element can be externally addressed. The evaluator is categorized as a subjective, due to the inherent influence of internal cognitive processes, prior experiences, cultural

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

background, and personal preferences on evaluative judgments. It is challenging to standardize without modifying the evaluator themselves or choosing evaluators who have had exactly the same experiences – a virtually impossible feat. In contrast, the evaluation process, evaluated product, and external factor are treated as objective factors, as their impact might be somewhat alleviated or organized through methodological and environmental design. This distinction does not imply that objective factors are bias-free, but rather that they are more manageable and might be easier to mitigate, compared to subjective evaluator-related factors. For example, several effects of objective factors in the **evaluation process**, such as the evaluation setting and presentation format, might potentially skew evaluations (Amabile, 1982). Specifically, the sequence of presentation of ideas may affect evaluators' perceptions; earlier or more distinctive designs might induce primacy or contrast effects (Amabile, 1982). Nevertheless, they might be alleviated by an appropriate setting to diminish their impact. In addition, **evaluated product** (ideas under evaluation), that have a high level of realism and detail, are often deemed more creative when compared to less visually compelling ideas, even though they may lack substantial conceptual originality – this phenomenon is termed the "beauty bias" in the evaluation of design creativity (Kudrowitz et al., 2012; Kwon & Kudrowitz, 2019). This bias might be mitigated by comparative and repeated evaluation. Furthermore, **external factor** also potentially affects design creativity evaluation, such as the physical environment (also possibly to be mitigated by a suitable setting). However, subjective factors may be challenging to mitigate or reduce. As mentioned, commonly used creativity evaluations require domain experts (**evaluator** in the framework) to evaluate design creativity. The experts bring their cultural background, prior experiences and preferences, which may influence their interpretation of the novelty and usefulness of ideas (Amabile, 1982; Christiaans, 2002; Niu & Sternberg, 2001).

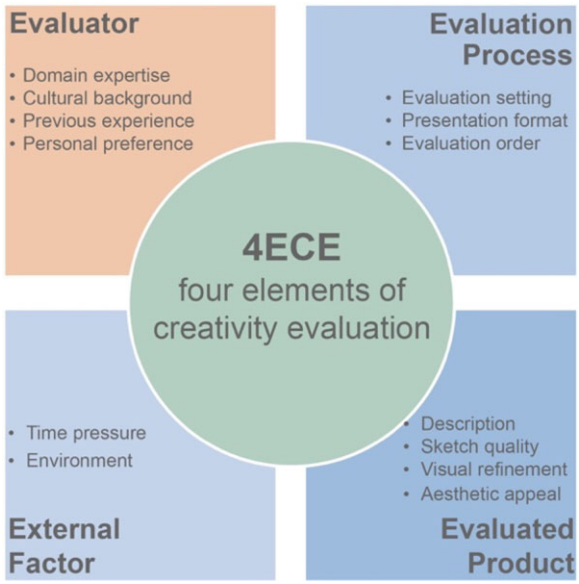


Figure 1. Factors influencing design creativity evaluation

2.4. Research gap and motivation

Despite extensive research on design creativity evaluation, existing techniques still encounter limitations, as explained above. Scoring-based methods, such as the CAT, provide valuable insights but can vary due to evaluator bias, preference, inconsistent selection, ranking order and the visual refinement of ideas (Figure 1).

To address these challenges, the current study introduces a ranking-based evaluation tool through triadic comparisons to mitigate objective factors in design creativity evaluation. Comparative methods have demonstrated greater consistency (Gill & Bramley, 2013; Pollitt, 2012; Tarricone & Newhouse, 2016), but have not yet been widely adopted in design research. By developing this tool and approach, the study aims to provide a reliable and effective approach for evaluating design creativity by mitigating several known objective influences in design creativity evaluation.

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

3. Tool development

3.1. Rethinking design creativity evaluation

The most recognized approach in the creativity field for evaluating creative output is absolute scoring. However, comparative judgement is grounded in cognitive psychology (Thurstone, 1974), with evidence showing that humans are more accurate in making relative than absolute judgements (Gill & Bramley, 2013; Pollitt, 2012; Tarricone & Newhouse, 2016). Comparative judgments were initially used to pair psychophysical stimuli, such as weight and height, and to measure values and attitudes (Thurstone, 1974). Its goal was to achieve immediate perception and minimize cognitive processing by repeatedly comparing pairs of items, objects, or performances (Tarricone & Newhouse, 2016). Researchers have argued that one reason comparative judgment is more accurate is that it limits evaluators to only two options: A or B, whereas absolute judgments have several levels to choose from (Gill & Bramley, 2013). Inspired by this evidence, we developed a novel tool and proposed an approach to evaluate design creativity through comparative evaluation. The high accuracy of comparative judgment may stem from raters' random selection in pair comparisons (Gill & Bramley, 2013); thus, we employed triadic comparisons instead of pair comparisons to eliminate random selection between two alternatives. Another reason for selecting triadic comparisons rather than four or five is that previous studies demonstrated that increasing the number of options inflates cognitive burden and difficulty of choice. The larger sets can reduce decision consistency and increase ambiguity in judgments (Cowan, 2010), making smaller comparison sets, such as triadic comparison, a well-established balance between contextual richness and cognitive manageability. However, it remains impractical to rely solely on triadic comparison, as a complete triadic comparison requires a large number of comparisons, many of which are redundant once transitive relations are established. For instance, when comparing A, B, and C, A surpasses B and C; in the comparison of D with E and F, D exceeds E and F; and when comparing B with D, B is superior to D. Consequently, all comparisons involving A and F, as well as, B and F, are unnecessary (e.g., comparisons between A and C, D, E, and F). Therefore, we attempted to integrate several algorithms and ranking methods into the tool to decrease the rounds of comparison. Finally, based on testing, we found integrated triadic comparisons with Tournament-based ranking to maintain a balance between precision and the number of comparison rounds. Tournament-based ranking exploits such transitivity to prune comparisons, reducing the required number extensively. For example, 24 ideas result in roughly 50–60 triadic comparisons using Tournament-based ranking, compared to 2024 rounds without it, balancing accuracy and feasibility.

3.2. Tool design and function

The evaluation tool was developed as an interactive, web-based system that combines intuitive interfaces for evaluators with a computer back-end for rating aggregation. The system facilitates the organized gathering, processing, and visualization of design creativity evaluation.

3.2.1. Rationale for comparative evaluation

As stated in previous section 3.1, we employed triadic comparison for design creativity evaluation. In each round, evaluators are presented with three images (ideas) simultaneously and rank them as best (+1), average (0), or worst (−1). In each ternary comparison, images are compared to each other to determine which ones are better than others. When image A is rated "best," B "average," and C "worst," the resulting preference chain is A → B → C. When evaluators continue to rate sets of images, these relationships are added to a directed graph. Each node represents an image, and each edge represents a preference (the "winner" points to the "loser"). This relational structure forms the basis for the tournament sort process, where comparative data are translated into a ranking that reflects the collective evaluation of ideas.

3.2.2. Ranking through Tournament Sort

We established a tournament-style ranking system, influenced by the Tournament Sort algorithm and pairwise comparison frameworks, to transform comparative evaluation outcomes into an organized ranking of design ideas. Tournaments are a commonly employed method for ranking (Orbán-Mihálykó et al., 2024; Sziklai et al., 2022; Zeng & Zeng, 2023). The algorithm builds directed edges (like A → B → C) and incorporates them into an expanding graph each time an evaluator completes a three-image comparison.

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

Once all images have been rated, the algorithm performs topological sorting to determine the final order of the images, prioritizing those with no incoming edges (i.e., images that are not outperformed by others). When the algorithm identifies a cycle, such as A beating B, B beating C, and C beating A, it resolves the confusion by considering both the number of ratings and the recency of those ratings. The sorting mechanism provides higher priority to images that have been rated more frequently. This approach enables the system to determine the optimal order when direct preference relationships result in circular dependencies. After resolving all cycles, the system examines adjacent images in the provisional order to see if they have a direct or indirect preference link. If no relationship exists, the pair is considered a tie. These tied images are flagged for additional comparison in subsequent rating rounds to improve the ranking.

This approach ensures that the ranking evolves iteratively: user feedback keeps strengthening the graph structure, resolving ties, and improving the ordering accuracy. As a result, the final ranking reflects both the frequency of wins and the relational strength among images, capturing the true comparative landscape of user preferences. The overall ranking workflow is illustrated in Figure 2, which shows how user ratings, graph construction, sorting, and tie-handling interact to produce a continuously improving ordered list.

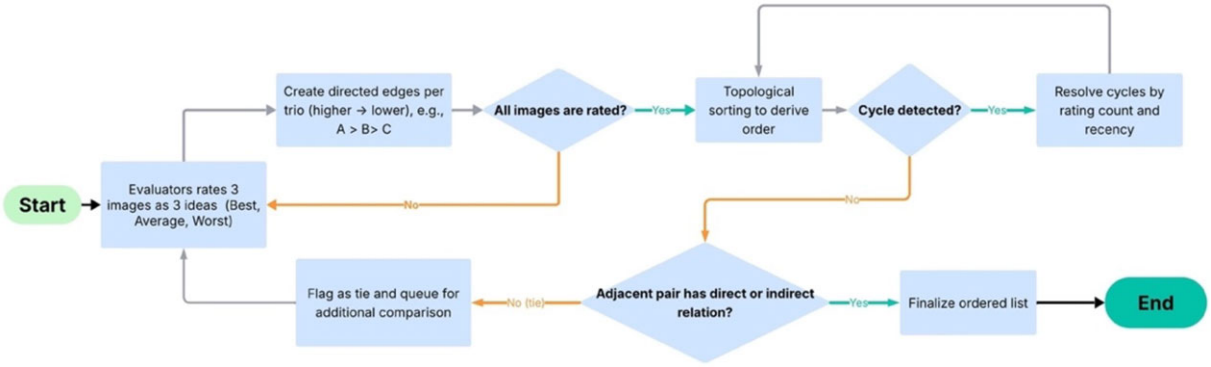


Figure 2. Workflow of the rating and ranking process. The system creates directed edges from ternary ratings, sorts them using topological sorting, resolves cycles using rating frequency and recency, and flags unresolved ties for future comparisons

3.3. Effectiveness of the tool

To test the tool's efficacy for ranking, we compiled a list of 24 images, and each image was labelled simply as "Idea 1," "Idea 2," and so on, up to "Idea 24," as the examples in Figure 3. Notably, the images are not actual ideas or sketches; they are simply blank pages labelled with numbers aiming to verify the effectiveness of the tool for ranking. For each comparison of three ideas, we assign a rating from 1 to 3 according to the idea's label. For example, if the screen displays three ideas, "Idea 5," "Idea 6," and "Idea 16," then we rank "Idea 5" as the best, "Idea 6" as the average, and "Idea 16" as the worst. When the ranking is from 1 to 24 according to the label of the ideas, it indicates that the tool effectively ranks based on the evaluators' input.



Figure 3. Part of the uploaded images

The procedure is outlined as follows (Figure 4):

1. Uploading images: We submitted 24 images (Figure 3) and then pressing "Start uploading," as illustrated in Figure 4. Upon the display indicating "Successfully uploaded," we could click "Start rating," which will instantly transition to the "Evaluating images" interface (Figure 4).
2. Evaluating images: we could view a screen displaying three ideas that we have to rank from 1 to 3 by clicking on the images (Figure 4). The first click denotes the most creative idea among the

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

three, the second click signifies the intermediate idea, and the third click represents the least creative idea, illustrated in Figure 4. To alter the ranking, we could click the icon in the upper right corner, which displays a delete symbol. To enhance the visibility of the idea, we can utilize the magnifying icon located in the bottom right corner. Upon finalizing the ranking and opting against any alterations, we proceed by clicking "Submit rating." Subsequently, we can rate another group of three ideas. After around 40 comparisons, the screen will instantly jump to "Viewing results."

3. Viewing results: This page displays the ranking of ideas, as anticipated, from "Idea 1" to "Idea 24". Furthermore, it displays the "Average Score" and "Rating Count" for each idea, as shown in the screenshot in the lower left corner of Figure 4.
4. Evaluation history: we can observe all comparisons and the time of each comparison (the screenshot in the lower right corner of Figure 4).

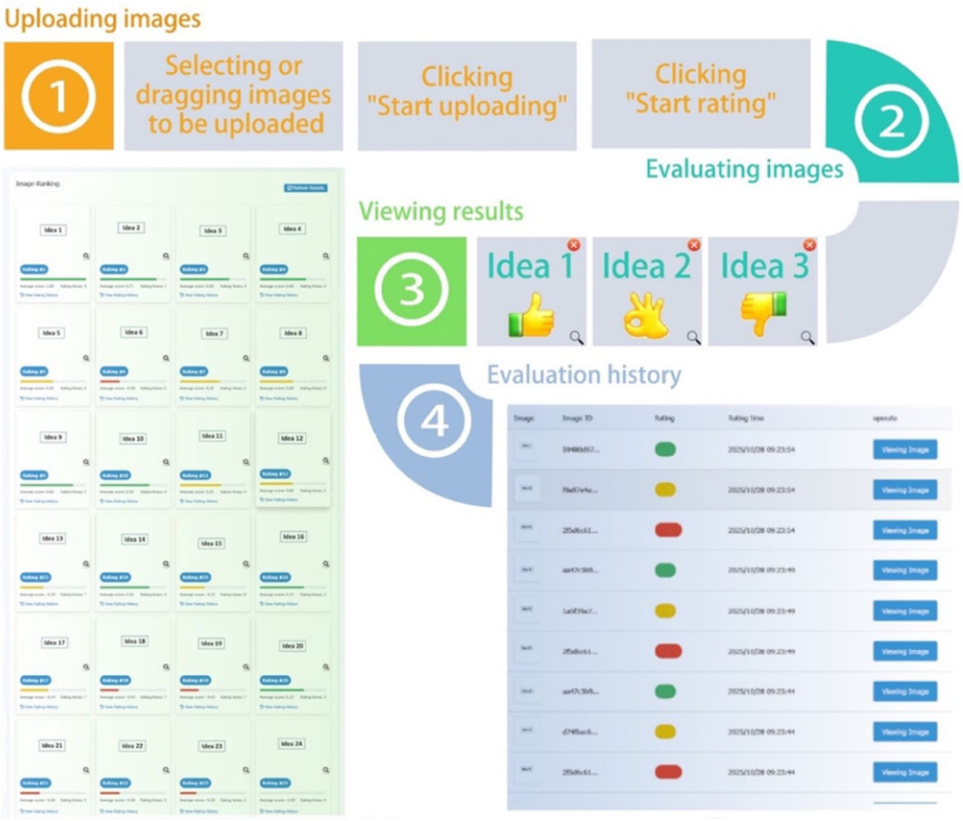


Figure 4. Procedure of tool utilization

3.4. Approach using the tool in design creativity evaluation

We recommend using the developed tool as a structured approach for evaluating design creativity for free (https://xy-image-analyse.onrender.com/). The approach is proposed as a theoretically grounded method suggestion for design creativity evaluation. While the tool was tested for its effectiveness (ranking accuracy), the suggested approach has not yet been empirically validated in real-world evaluation settings. The following procedure outlines the suggested steps for the approach integrated with the tool:

- **Preparation.** Before conducting the evaluation, ensure that experts (evaluators) have the related domain expertise to rate design creativity. In addition, they have sufficient time and are not under pressure from other tasks. Adequate focus and mental readiness are essential for consistent and thoughtful evaluation.
- **Familiarization with ideas.** After collecting all design ideas to be evaluated, invite experts to review them in advance, and they are encouraged to speak out and share their understanding of each idea. During this stage, evaluators should briefly explain their understanding of each idea and identify any ideas they find unclear. Ideas that cannot be understood should be excluded from further evaluation to maintain fairness and consistency.

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

- **Evaluation using the tool.** Once familiarized, evaluators use the tool to conduct triadic comparisons and generate rankings. During the evaluation, they are encouraged to speak loudly about the reasons for their ranking. Evaluations can be based on one or multiple criteria, such as creativity, novelty, and usefulness. When multiple criteria are evaluated, evaluators should complete separate rounds of comparisons for each criterion. The second and subsequent rounds typically proceed more quickly, as evaluators are to be already familiar with the idea set.
- **Weighting and scoring.** After obtaining the ranked results, a weighted scoring scheme can be applied to quantify the outcomes. For example, for only evaluating creativity in three levels, following the study of Dean et al. (2006), ideas in the top 10% of the ranking can be assigned a high score (3), representing highly creative ideas; the bottom 10% can be assigned a low score (1), indicating low creativity; and the remaining 80% can be assigned a moderate score (2). This weighted approach translates relative rankings into quantitative data for further analysis.

4. Discussion and implication

In the section, we present our insights related to the design creativity field and interpret how the developed tool and approach might be helpful to mitigate several objective influences on design creativity evaluation. First, we identify the factors that might be addressed through the tool and approach, demonstrating its potential to mitigate influences on design creativity evaluation. We also discuss factors that remain challenging to mitigate, explaining the inherent limitations of design creativity evaluation. Furthermore, we highlight the broader implications of the tool, suggesting its potential applicability across diverse fields, such as academics, education and industry.

4.1. Mitigated influences in design creativity evaluation

As discussed earlier, several **objective** influences in design creativity evaluation might be mitigated through the application of the developed tool and approach, particularly those related to **evaluated product** and **evaluation process** factors. Because the **external factor**, while not directly addressed by the presented tool and approach, they may be alleviated through appropriate configurations, by asking the evaluators' preferred physical environment prior to the evaluation and adjusting it accordingly.

One common **evaluation process**-related influence concerns the evaluation order of ideas. For example, when a creative idea appears early in the evaluation sequence and similar ideas appear later, the first idea may receive a higher score simply because it is perceived as novel at first glance. In contrast, subsequent similar ideas may be rated lower, as evaluators have already seen a comparable idea and thus perceive it as less creative. The developed approach helps mitigate this effect in several ways. First, evaluators are required to review all ideas before beginning the evaluation, allowing them to form an initial understanding of the entire idea pool. Second, the tool uses a comparative evaluation approach with randomized pairings, ensuring that similar ideas are eventually compared against each other rather than being influenced by their presentation order. Through this process, the potentially reduces the influence caused by the sequence of idea presentation and supports a fairer and more consistent evaluation of design creativity.

Regarding **evaluated product**-related factors, such as the sketch quality or visual presentation of ideas that influences perceived creativity (Kudrowitz et al., 2012), the developed evaluation tool might also help mitigate these effects by emphasizing the evaluation criteria used. For instance, evaluators are explicitly reminded to focus on the designated metric of design creativity, such as novelty or usefulness, rather than drawing style or visual appeal. Evaluators are also encouraged to explain their reasoning aloud during the evaluation, explaining why one idea is considered more creative/novel/useful than another. Moreover, when evaluators hesitate or make inconsistent rankings between similar ideas, the system automatically provides higher priority to ideas that have been rated more frequently. In this way, the tool ascertains the ideal sequence when direct preference interactions create circular dependencies. Through these procedures, the tool promotes reflective evaluation and reduces the unintentional influences that may arise from differences in idea presentation quality.

4.2. Challenges in mitigating influences on design creativity evaluation

However, several **subjective influences** remain difficult to eliminate (e.g., **evaluator** in the **4ECE**), such as those related to evaluators' prior experiences and personal preferences. For example, an idea may be rated as highly creative by evaluators who have never encountered a similar idea, while another evaluator

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

with prior exposure to comparable ideas may consider it less creative. Such differences can occur even among domain experts, as creativity evaluation is often shaped by individual knowledge, backgrounds and experiences.

Additionally, personal preferences may also affect evaluations. Nevertheless, the use of the developed tool might help to reduce (though not completely remove) these influences. Because the tool employs a comparative evaluation method, evaluators are encouraged to focus on the defined criteria (e.g., creativity, novelty, usefulness) rather than their subjective impressions or preferences. In cases where two ideas exhibit similar levels of creativity, an evaluator's preference might still influence the ranking slightly; however, during the weighting stage, these ideas would likely fall within the same scoring category, thereby minimizing the overall impact of individual preference on the final results. The tool may also alleviate evaluator fatigue. In contrast to conventional approaches that necessitate evaluators to score numerous ideas, this tool requires a straightforward, relative ranking among three ideas at a time. This comparative evaluation reduces cognitive strain and maintains evaluators' focus and precision during extended sessions.

4.3. Implication of the developed tool

The created tool and approach for design creativity evaluation have substantial significance for academic, educational and industrial settings. The tool facilitates an organized, comparative, and transparent review process, enhancing the rigor, consistency, and reflection of diverse expert viewpoints in design creativity evaluation.

In **academics**, the proposed tool and approach advance design creativity evaluation by providing a systematic and efficient alternative to the traditional CAT, established nearly half a century ago and still the prevailing method for evaluating creativity. Despite decades of research in design creativity, evaluations still rely heavily on CAT with two domain experts, facing subjective and objective influences. In contrast, our comparative evaluation tool reframes creativity evaluation as a structured and data-driven process. Through triadic comparisons and rank aggregation, it captures evaluators' relative judgments more consistently while reducing cognitive load and providing more rigor, consistency, and reflection of evaluation. The tool's digital implementation further supports reproducibility and scalability, representing a modern enhancement and complement to the long-standing CAT method.

In **design education**, the tool can function as a pedagogical resource to assist students in comprehending the perception and design creativity evaluation. In addition, it could help educators fairly evaluate students' ideas and products. Furthermore, engaging students in comparison evaluations allows them to understand the variability of creative ideas in terms of novelty and usefulness. The tool may further serve in peer evaluation activities, fostering debate and reflection on creative processes and results.

The tool aids design teams, innovation managers, and research and design departments in evaluating ideas throughout the first phases of product development in **industry**. Conventional brainstorming or ideation sessions frequently generate several ideas; however, evaluating the creativity of ideas can be difficult, as it is variable and subjective. The tool might reduce uncertainty by offering a consistent way to rate ideas, ensuring that the evaluation results come from a group of experts instead of personal opinions. This might improve decision-making efficiency and result in more rigorous idea selection procedures. Moreover, the tool's adaptability for various evaluation criteria renders it suitable for a wide range of fields, including product design, service innovation, digital experience design, and engineering solutions.

5. Limitations and future work

Our current study has several limitations. First, due to the number of calculations and comparisons required, the tool is best suited for evaluating a relatively small number of ideas (fewer than 50). For example, in our test with 24 ideas, approximately 40 comparison rounds and around 5 minutes were needed, which was manageable. However, in practical implication, evaluating 24 ideas may require approximately 15 minutes, involving 38 to 55 comparative rounds for a single metric. The number of comparison rounds is significantly influenced by the evaluators' inconsistent choices, prompting the system to automatically generate additional rounds to guarantee ranking accuracy while prioritizing ideas that have received more frequent and recent ratings. Second, while we successfully designed and tested the functionality and effectiveness of the tool, it has not yet been validated in a real-world study. For the proposed approach, which integrates the tool as we suggested, it is also not verified by a real-world setting. Future research will focus on applying the tool and approach in practical design creativity

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

evaluation by recruiting evaluators from the design field. This will allow us to assess its usability, reliability, and potential biases in real evaluation scenarios. By conducting such empirical studies, we aim to identify both the strengths and limitations of the tool and approach, ultimately refining its design and expanding its applicability. We believe that this tool represents a novel and rigorous approach to evaluating design creativity, offering greater consistency and objectivity in the evaluation process.

6. Conclusions

Due to the challenges of design creativity evaluation, we initially developed a framework, named the **4ECE**, which summarized four influences (evaluator, evaluation process, evaluated product, and external factor) on the evaluation of creativity and divided them into subjective and objective factors. In addition, inspired by the comparative judgement, we created an innovative **tool** that offers a more stringent and objective framework for evaluating design creativity. The tool relies on **comparative evaluation** rather than previous absolute evaluation, which is believed to be more objective and to reduce several recognized factors that affect the evaluation of design creativity. Moreover, we assessed the efficacy of the tool and suggested **a comprehensive approach** on its use. Additionally, we discussed the advantages of the tool for evaluating design creativity, which may help reduce certain objective factors that influence design creativity evaluation, while also highlighting the inevitable subjective influences involved in evaluating design creativity. Finally, we emphasized the significance of the tool in academics, education and industry. The established tool and approach might improve research and practice in evaluating design creativity by linking **subjective human evaluation** with **systematic assessment**, thereby offering a foundation for evidence-based improvements in design creativity evaluation.

References

Abraham, A. (2022). Creativity or creativities? Why context matters. *Design Studies*, 78, 101060. https://doi.org/10.1016/j.destud.2021.101060

Amabile, T. M. (1982). Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013.

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65–77). IGI Global.

Bastian, A., Luo, F., Xia, T., & Albers, A. (2025). Improving cultural understanding and sparking creativity through the application of the CSS in an intercultural and primarily Chinese research team. *Design Science*, 11, e22. https://doi.org/10.1017/dsj.2025.10013

Borgianni, Y., Maccioni, L., Fiorineschi, L., & Rotini, F. (2020). Forms of stimuli and their effects on idea generation in terms of creativity metrics and non-obviousness. *International Journal of Design Creativity and Innovation*, 8(3), 147–164. https://doi.org/10.1080/21650349.2020.1766379

Casakin, H., & Georgiev, G. V. (2021). Design creativity and the semantic analysis of conversations in the design studio. *International Journal of Design Creativity and Innovation*, 9(1), 61–77. https://doi.org/10.1080/21650349.2020.1838331

Çelik, T. (2025). The effect of artificial intelligence on creativity in conceptual design in architectural education: The motion of biomimetics and futurism. *AI & SOCIETY*, 40(4), 2717–2727. https://doi.org/10.1007/s00146-024-02042-7

Chakrabarti, A. (2009). Design Creativity Research. In N. R. S. Raghavan & J. A. Cafeo (Eds), *Product Research: The Art and Science Behind Successful Product Launches* (pp. 17–39). Springer Netherlands. https://doi.org/10.1007/978-90-481-2860-0_2

Christiaans, H. H. C. M. (2002). Creativity as a Design Criterion. *Creativity Research Journal*, 14(1), 41–54. https://doi.org/10.1207/S15326934CRJ1401_4

Cowan, N. (2010). The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, 19(1), 51–57. https://doi.org/10.1177/0963721409359277

Dean, D., Hender, J., Henley Management College, Rodgers, T., Consultant College Station, Texas, Santanen, E., & Bucknell University. (2006). Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7(10), 646–699. https://doi.org/10.17705/1jais.00106

Georgiev, G. V., & Casakin, H. (2019). Semantic Measures for Enhancing Creativity in Design Education. *Proceedings of the Design Society: International Conference on Engineering Design*, 1(1), 369–378. https://doi.org/10.1017/dsi.2019.40

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice*, 20(3), 308–324. https://doi.org/10.1080/0969594X.2013.779229

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press

Gonçalves, M., Cardoso, C., & Badke-Schaub, P. (2016). Inspiration choices that matter: The selection of external stimuli during ideation. *Design Science*, 2, e10. https://doi.org/10.1017/dsj.2016.10
Gonçalves, M., & Cash, P. (2021). The life cycle of creative ideas: Towards a dual-process theory of ideation. *Design Studies*, 72, 100988. https://doi.org/10.1016/j.destud.2020.100988
Gong, Z. (2024). Understanding cultural influence on creativity in ideation [University of Oulu]. https://urn.fi/URN:NBN:fi:oulu-202312123674
Gong, Z., Gonçalves, M., Nanjappan, V., & Georgiev, G. V. (2024). Priming uncertainty avoidance values: Influence of virtual reality stimuli on design creativity in ideation. *Computers in Human Behavior*, 158, 108257. https://doi.org/10.1016/j.chb.2024.108257
Gong, Z., Lee, L.-H., Soomro, S. A., Nanjappan, V., & Georgiev, G. V. (2022). A Systematic Review of Virtual Brainstorming from the Perspective of Creativity: Affordances, Framework, and Outlook. *Digital Creativity*, 33(2), 96–127. https://doi.org/10.1080/14626268.2022.2064879
Jagtap, S., Larsson, A., Hiort, V., Olander, E., & Warell, A. (2015). Interdependency between average novelty, individual average novelty, and variety. *International Journal of Design Creativity and Innovation*, 3(1), 43–60. https://doi.org/10.1080/21650349.2014.887987
Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20(2), 171–178. https://doi.org/10.1080/10400410802059929
Kudrowitz, B., Te, P., & Wallace, D. (2012). The influence of sketch quality on perception of product-idea creativity. *AI EDAM*, 26(3), 267–279. https://doi.org/10.1017/S0890060412000145
Kwon, J., & Kudrowitz, B. (2019, November 25). The Sketch Quality Bias: Evaluating Descriptions of Product Ideas With and Without Visuals. ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. https://doi.org/10.1115/DETC2019-97232
Niu, W., & Sternberg, R. J. (2001). Cultural influences on artistic creativity and its evaluation. *International Journal of Psychology*, 36(4), 225–241. https://doi.org/10.1080/00207590143000036
Orbán-Mihálykó, É., Mihálykó, C., & Gyarmati, L. (2024). Evaluating the capacity of paired comparison methods to aggregate rankings of separate groups. *Central European Journal of Operations Research*, 32(1), 109–129. https://doi.org/10.1007/s10100-023-00839-3
Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x
Rhodes, M. (1961). An Analysis of Creativity. *The Phi Delta Kappan*, 42(7), 305–310.
Sääksjärvi, M., & Gonçalves, M. (2018). Creativity and meaning: Including meaning as a component of creative solutions. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 32(4), 365–379. https://doi.org/10.1017/S0890060418000112
Schwarz, N., & Sudman, S. (2012). *Context Effects in Social and Psychological Research*. Springer Science & Business Media.
Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, 24(2), 111–134. https://doi.org/10.1016/S0142-694X(02)00034-0
Sziklai, B. R., Biró, P., & Csató, L. (2022). The efficacy of tournament designs. *Computers & Operations Research*, 144, 105821. https://doi.org/10.1016/j.cor.2022.105821
Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, 13(1), 16. https://doi.org/10.1186/s41239-016-0018-x
Thurstone, L. L. (1974). A Law of Comparative Judgment. In *Scaling*. Routledge.
Verhaegen, P.-A., Vandevenne, D., Peeters, J., & Duflou, J. R. (2013). Refinements to the variety metric for idea evaluation. *Design Studies*, 34(2), 243–263. https://doi.org/10.1016/j.destud.2012.08.003
Zeng, E. Z., & Zeng, Y. (2023). Beyond the Scoreboard: Advancing Fairness in Athlete Selection with Simulation-Based Tournament Strategies. *Journal of Computational Vision and Imaging Systems*, 9(1).

https://doi.org/10.1017/pds.2026.10460 Published online by Cambridge University Press